TOWARDS MULTILINGUAL INTEROPERABILITY IN AUTOMATIC SPEECH RECOGNITION

Martine Adda-Decker
Spoken Language Processing Group, LIMSI-CNRS
Bât 508, BP 133, 91403 Orsay, Cedex, France
madda@limsi.fr http://www.limsi.fr/TLP

ABSTRACT

In this communication, we address multilingual interoperability aspects in speech recognition. After giving a tentative definition of multilingual interoperability, we discuss speech recognition components and their language-specific aspects. We give a sample overview of past multilingual speech recognition research and development across different speaking styles (read, prepared and conversational). The problem of adaptation to new languages is addressed. Language-independent and cross-language techniques for acoustic modeling provide a means to port recognition systems to new languages without language specific acoustic data. Pronunciation lexica and text material appear to be the most crucial language-dependent resources for porting. Fast porting being a step towards multilingual interoperability the ongoing efforts of producing multilingual pronunciation lexica and collecting multilingual text corpora should be extended to the largest possible number of written languages.

1. INTRODUCTION

The important progress achieved in speech recognition these last decades has led to successful demos using speech technology. Demos raise expectations when shown to potential users, but yet only few systems are ready for operational use. In a multilingual environment, where potential users have distinct native languages, speech recognition systems have to deal with these different languages or with non-native speaker accents, if a common language is shared. Multilingual environments are common in international communication contexts, which may be political, military, scientific, commercial or tourist contexts. The development of multilingual recognition and spoken dialog systems is hence an important research issue, opening a large spectrum of potential applications. To increase the usability of a prototype system the problems of multilingual and non-native speech have to be addressed efficiently.

Speech recognizers are still very sensitive to non-native speech input or more generally to any kind of condition mismatch. Porting a given system to a new language requires often a significant part of language specific knowledge and resources before achieving viable recognition results. Multilingual corpora have been gathered for language identification and multi-lingual recognition research (OGI-TS, LDC CALLHOME, GLOBALPHONE...). Research and development in multilingual recognition has been widely supported by the European communities (EC) and the Defense Advanced Research Project Agency (DARPA) [39, 5, 12, 40, 14, 43].

In this contribution we address issues of multilinguality and multilingual interoperability in speech recognition.

Using a standard recognizer architecture based an acoustic HMM phone models, pronunciation dictionaries and word N-gram language models, the language-specific aspects of each component are discussed. Many observations are gathered from our experience at LIMSI in developing multilingual speech recognizers [35, 54, 2, 1, 4]. We will then focus on multilingual recognition systems. Without attempting to be exhaustive we try to give an overview of some representative research actions in multilingual and cross-lingual speech recognition.

2. MULTILINGUALITY AND MULTILINGUAL INTEROPERABILITY

There exist about 3000 different spoken languages without accounting for dialects, at the end of this millennium [38]. According to this author only several 100 languages have also a significant written language production for which current speech recognition systems (speech to text systems) are applicable. Studies in automatic speech recognition (ASR) are presently limited to about 20 languages, comprising English, Arabic, Chinese, Japanese, Spanish, French, German, Italian, Portuguese, Greek, Swedish, Danish, Dutch...

Interoperability is a term which is widely used in product marketing descriptions: products achieve interoperability with other products either by adhering to published interface standards (example: the WEB with standards such as TCP/IP, HTTP, HTML) or by making use of a "broker" of services that can convert one product's interface into another product's interface on the fly (example: common object request broker architecture CORBA). Interoperability becomes a quality of increasing importance for information technology products, and naturally, the demand for interoperability of speech technology products arises. Voice over IP (VoIP) protocols have already evolved into world-wide standards (IETF's SIP, ITU,s H.323) to support the emerging voice, data and video services of the next millennium.

For speech recognition systems the term of interoperability is not yet commonly used in the corresponding researcher community. Nonetheless many past or present research actions aim at defining standards for text and speech processing (e.g. the EC EAGLES project on language engineering standards [26]), at developing multilingual resources ([51, 45, 15, 12, 5]), at installing

multilingual recognizer evaluations (e.g. the EC SQALE project on multilingual speech recognition evaluation, the DARPA Hub5 program on conversational multilingual speech), and at achieving larger robustness across varying experimental conditions (e.g. the DARPA Hub3 program and Hub4 broadcast news transcriptions). Research towards better multilingual interoperability is supported and fostered by national and international institutions: EC (European Commission), NSF (National Science Foundation), DARPA...

Multilingual interoperability which is the topic of this workshop deals with the problem of designing speech products which are operative in a multilingual context and/or easily portable to new languages. The development of multilingual corpora and resources can be considered as a milestone on the way to multilingual interoperability. Developing such resources however is time-consuming, expensive and their reusability is not always ensured, when moving to new application domains. Important related research areas concern cross-domain portability. Research directions towards more language-independent approaches for speech recognition are also being investigated[47, 32, 31] especially for acoustic modeling.

3. SPEECH RECOGNITION

We briefly review the main components of the recognizer in a statistical approach commonly used for LVSR (*Large Vocabulary Speech Recognition*) [6], [27], [53] and discuss to what extend these components are language-specific. The speech recognizer has to determine the most probable word sequence $\widehat{w_1^N}$ given the acoustic input x_1^T:

$$\widehat{w_1^N} = \arg\max_{\{w_1^n\}} \Pr(w_1^n)\Pr(x_1^T|w_1^n)$$

where w_1^n is a sequence of n words each in the lexicon, n being a positive integer. The acoustic input x_1^T is a feature stream, chosen so as to reduce model complexity while trying to keep the relevant information (i.e. the linguistic information for the speech recognition problem). While the use of language-dependent acoustic features has been investigated (see dedicated session of ICSLP'98) acoustic parameter extraction can be considered as mostly language-independent.

$\Pr(w)$ is to be provided by a language model, and $\Pr(x|w)$ by an acoustic model. The recognition decision is taken as a joint optimization of two terms: $\Pr(w)$, the a priori probability of a word or a word sequence as given by the language model and $\Pr(x|w)$ the conditional probability of the signal corresponding to the word sequence, given by the acoustic model. The output $\widehat{w_1^N}$ is a sequence of items from the vocabulary $\{w_i\}$. Pronounced items which are not in the lexicon (referred to as out-of-vocabulary words or OOVs) are necessarily missing in the recognizer's output, and thus misrecognized. Hence the motivation for maximizing lexical coverage by appropriate definition and selection of the lexical items during training.

- **the acoustic model** $\Pr(x|w)$
  Acoustic units generally correspond to subword units which when compared with word models, reduce the number of parameters, enable cross word modeling and porting to new vocabularies in a monolingual context. For Hidden Markov Model (HMM) based systems acoustic modeling most commonly makes use of context-dependent (CD) phone units.[1] $\Pr(x|w)$ is then obtained via a pronunciation lexicon, where each word w_i is described as a sequence of the appropriate phones:

$$\Phi(w_i) = \phi_1^i \oplus \phi_2^i \oplus \ldots \phi_m^i$$

$$\Pr(x|w_i) \equiv \Pr(x|\Phi(w_i)) = \Pr(x|\phi_1^i \oplus \phi_2^i \oplus \ldots \phi_m^i)$$

  Consistent use of the different phone symbols in the lexicon is probably the most important requirement in pronunciation generation. CD models allow for implicit coarticulation modeling within the acoustic model. Coarticulation due to the surrounding phones necessarily occurs for all languages and hence context modeling should be an effective approach for any language. As CI models merge all different coarticulation effects within the same model, they are more robust as compared to CD models. Separating coarticulation effects using an increasing number of contexts results in a more accurate representation of the acoustic patterns. CD models, accounting for the phonotactic constraints of the language, are hence more language-specific than CI models. Concerning the acoustic phone models (CI or CD) we have to be aware that they always best model the most frequently observed coarticulation effects of the training data. For training corpora with a low lexical variety, CI phone models tend to become word-dependent with possibly poor generalization abilities, both intra and inter language.
  Language-dependent CI models (and even recently context-dependent phone models [31]) have been experimented with for porting a recognizer to new languages.
  To overcome the problem of unobserved sounds when porting acoustic models to a new language, studies aiming at developing multilingual or language-independent acoustic phone models are undertaken both for speech recognition and language identification. Recent researches on language-independent acoustic phone models and cross-language adaptation can be found in [47, 32, 31, 16]. These studies tend to demonstrate the viability of a language-independent acoustic modeling approach. Whereas it is important to be able to bootstrap a recognizer for a new language without prior acoustic models of that language, most researchers tend nonetheless to conclude that using a small amount of language-specific acoustic data either to train language-dependent models or to carry out a language-dependent adaptation, rapidly outperforms foreign language data. MLLR [37] and MAP adaptation techniques are used for adapting cross-lingual or multilingual acoustic models to the new language.
- **the language model** $\Pr(w)$
  Language models are used to model regularities in natural language, and can therefore be used in speech recognition to predict probable word sequences during decoding. The most popular methods, such as statistical n-gram models, attempt to capture the syntactic and semantic constraints

[1] In some real-time systems context-independent (CI) phone units may be used in order to reduce the computation time and search space.

by estimating the frequencies of sequences of n words. The lexical unit, w_i, can be considered the basic observation for statistical language models. The extraction of w_i units from text sources can be more or less straightforward depending on the language (e.g. easy for English or French, difficult in Japanese: no spacing between words)

Given a fixed amount of training data, less reliable language models (LMs) are usually obtained for highly inflected languages (with large lexical variety) than for less inflected languages. The same observation can be made for agglutinative languages. In the latter case decompounding could be applied for lexical unit definition. Tokenizations or text normalizations aimed at reducing lexical variety include some language-independent and a variable amount of more or less complex language-dependent processing [1, 24].

The effectiveness of N-gram LMs for a given language also depends on the validity of the approximation of capturing the language structure within sequences of N words. We know that the validity of this approximation is strongly language-dependent, and hence the N-gram modeling approach will not give the same benefit to speech recognition systems for all languages, even if no limit on available training data were imposed.

- **the decoder** $\arg\max_{\{w_1^n\}}$
  The search space to be explored by the decoder is related to the lexicon size and the language model (LM) complexity. For a bigram LM the search space is proportional to the lexicon size. Pronunciation variants introduce additional entries in the search space. Computational requirements can be controlled by limiting LM size, lexicon size and pronunciation variants.

A speech recognizer should meet the following requirements to guarantee good performance. The vocabulary, the acoustic and language models have to achieve good coverage during the system's operating conditions. The vocabulary should thus contain all or most words likely to appear during operation. This means that the out of vocabulary (OOV) word rate should be minimal. Acoustic models should be able to accurately model the vocabulary words. Context-dependent models allowing for a high coverage of the vocabulary are likely to produce better results, than context-independent models or contextual models which are seldom observed during operation. Similarly language models should produce low perplexity during operation. The same criteria have to be met by multilingual systems.

4. MULTILINGUAL SPEECH RECOGNITION

Ideally a multilingual speech recognizer is able to transcribe speech from different languages, thus identifying both the language used and the word sequence uttered by the speaker. Whereas language and word string can be identified in parallel (multi-lingual recognizer), a more effective way, at least for now, is to prior identify the language using a language identification system on homogeneous acoustic segments, and then decode the word string with the appropriate language-dependent recognizer.

Existing systems have been developed for specific domains and a restricted number of languages, requiring large amounts of annotated language-specific corpora. Without trying to be exhaustive, we can cite some examples of multilingual recognizer developments: the LE-Sqale project on read speech LVSR in English, German and French [35, 54], the DARPA Hub5 program on conversational and multilingual speech LVCSR (*Large Vocabulary Conversational Speech Recognition*) over telephone [9, 12] using SWITCHBOARD and CALLHOME corpora.

4.1. Multilingual LVSR using read speech

The aim of the EC SQALE project (Speech recognizer Quality Assessment for Linguistic Engineering) was to experiment with installing in Europe a multilingual evaluation paradigm for the assessment of large vocabulary, continuous speech recognition systems (LVSR) to assess language-dependent issues in multilingual recognizer evaluation. This project, running from 1993 to 1995 gathered CUED Cambridge (UK), Philips Aachen (Germany), LIMSI Paris (France) and TNO Soesterberg (Netherlands).

In the SQALE project, the same system is being evaluated on comparable tasks in different languages (American English, British English, French and German) to determine cross-lingual differences. The recognizer makes use of phone-based continuous density HMM for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling. The system has been evaluated on a dictation task developed with read, newspaper-based corpora, the ARPA *Wall Street Journal* corpus of American English, the WSJCAM0 corpus for British English, the BREF-*Le Monde* corpus of French and the PHONDAT-*Frankfurter Rundschau* corpus for German. Experimental results under closely matched conditions are reported. The average word accuracy across all 4 languages is about 85%, obtained for a 20k vocabulary open test (65k open test for German) on a multilingual test set where the OOV rates are kept comparable across languages (about 2% OOVs) Trigram LMs and context-dependent acoustic models were used (about 800 CD models for French and more than 2500 tied-state CD models for English and German). A similar recognizer was developed in Japan [42] using 180M business newspapers. With a 7k vocabulary and an appropriate 7k test set without OOV words, an 80% word accuracy rate is achieved using a bigram LM and about 700 CD models.

In Tab. 1, lexical variety across different languages was investigated for comparable amounts of text corpora[2]. Coverage figures of Japanese reported in [42] are very close to those obtained for Italian. Whereas English achieves the highest lexical coverage (close to 100% for a 65k vocabulary, German has the highest OOV rate of about 5%. For a given speech technology (e.g. a 65k system) better results can thus be expected for English than for German. In German, a major obstacle to high lexical coverage arises from inflected forms and word compounding

[2]The newspaper text corpora compared are the *Wall Street Journal* (American English), *Le Monde* (French), *Frankfurter Rundschau* (German) from the ACL-ECI cdrom, *Il Sole 24 Ore* (Italian), and Nikkei (Japanese).

language	English	Italian	French	German	Japanese
corpus	*WSJ*	*Sole 24*	*Le Monde*	*FR*	*Nikkei*
#words	37.2M	25.7M	37.7M	36M	180M
#distinct	165k	200k	280k	650k	623k
5k cover.	90.6	88.3	85.2	82.9	88.0
20k cover.%	97.5	96.3	94.7	90.0	96.2
65k cover.%	99.6	99.0	98.3	95.1	99.2
20k-OOV%	2.5	3.7	5.3	10.0	3.8
65k-OOV%	0.4	1.0	1.7	4.9	0.8

Table 1: Comparison of *WSJ*, *Il Sole 24 Ore*, *Le Monde* , *Frankfurter Rundschau* and *Nikkei* text corpora in terms of number of distinct words and lexical coverage of the text data for different lexicon sizes. OOV rates are shown for 20k and 65k lexica.

for which morphological decomposition could be effectively applied.

More recently within the German GLOBALPHONE project a multilingual read speech database comprising 15 languages (Arabic, Chinese, Croatian, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil and Turkish) has been collected. Using these data the University of Karlsruhe is working on a multilingual LVSR system [47]. Their research efforts focus on multilingual acoustic modeling and fast bootstrapping of acoustic models for new languages. Speech recognition results have been obtained for 6 languages (word error rates ranging from 10% to near 50%) using 10k vocabularies. Closed test sets have been used by adding missing words in the vocabularies and assigning a low probability to the corresponding monograms in the LM. The multilingual text material is yet too limited for reliable language model estimation.

Experiments in multilingual read speech recognition indicate that good performances can be achieved across languages, provided that sufficient training material is available (10-100 hours of speech, 50-200M of words).

4.2. Multilingual LVCSR using conversational speech

The CALLHOME program [14] (part of the DARPA Hub5 program) was initiated in the US in 1995 in order to study conversational speech between family members over long-distance telephone in a multilingual context. Corpora were recorded in English, Mandarin, Japanese and Spanish (with a variety of dialects) during 1995, Arabic (colloquial Egyptian) and German during 1996. LDC provided the multilingual data to participants. Word error rate results reported in 1997 range from about 40% for English to around 60% for Spanish, Arabic, Mandarin and German. As stated by G. Zavaliakos [55], work on CALLHOME Corpora has verified that current technology is largely language independent. The better results obtained in English can be related to relatively more training data available in this language and maybe a longer and more reliable expertise in English system development. Nonetheless word error rates remain high across the different languages, significantly higher than those reported for read or prepared broadcast speech (around 20% word error rates, Hub4 DARPA program). To measure the impact of mere speaking style on recognition results, by controlling speaker, channel and LM effects, an interesting experiment was carried out at SRI as reported in [14]. Conversational speech was recorded and then transcribed. The same speakers were then invited to read the transcriptions, imitating spontaneous style and a second time in pure read style. Word error rates of about 50% for the true conversational style, drop to about 40% for the false spontaneous elocution, and to around 30% for the read version. Conversational speech doesn't fit the spoken language modeling assumptions as well as read speech (see section 3.). This is particularly true for the articulated phone sequence assumption of the pronunciation lexicon.

Results are consistently disappointing across languages for conversational speech. Whereas read or broadcast speech can be considered as normative to be understood by a large audience, familiar conversational speech spreads a larger variety of individual speaking styles. This may explain the discrepancy observed between performance in read and conversational speech. For the CALLHOME languages about 15 hours of acoustic training data and about 150k words for language model estimation were available. Vocabulary sizes ranged from about 10k to about 20k [12]. Experience taken from conversational speech in English (using Switchboard) shows that significant error reduction (i.e. better conversational speech modeling) can be achieved when moving from 15 to 150 hours of speech and from 150k to 2M words.

4.3. Multilingual Broadcast Transcriptions

The DARPA-Hub4 program, introduced in 1995, concerns broadcast news transcription.

Within the Broadcast transcription program, data collection and corpus design have become more efficient, as large amounts of news are constantly available. Corpus transcription and annotation standards [10] have been developed. Annotated corpora are easily created using freeware transcribing tools [7]. Human broadcast transcription/annotation can range from 10-50 times real-time.

Whereas the main effort is centered on English sources, non English (multilingual) evaluations have been carried out for Spanish [25] and Chinese systems [56], demonstrating the feasibility for other languages. English best results are below 20% word error rate. Error rates on non-native speech (F5 condition [48]) are higher for the corresponding native condition (F0),

but the F5 proportion remains low in the overall test sets.

Automatically generated broadcast news transcripts can be used for indexing or document retrieval tasks (NIST SDR program). These research areas go in the direction of speech understanding. The benefits of the Broadcast news task on speech recognition technology progress is discussed in [33].

In Europe the EC is also sponsoring research on multilingual broadcast transcriptions. As an example we can cite the LE4-OLIVE project launched in 1998, which aims to support automated indexing of video material by use of human language technologies and in particular multilingual speech recognition. The prime interest of the OLIVE users is to obtain an efficient, detailed and direct access to their video archives. The users in the OLIVE consortium are two television stations, comprising ARTE (Strasbourg, France) and TROS (Hilversum, Netherlands), as well as the French national audio-video archive, INA/Inatheque in Paris, France, and NOB, a large service provider for broadcasting and TV productions (Hilversum, Netherlands). Technology development and system implementation involve: TNO-TPD (Delft), the project co-ordinator supplying the core indexing and retrieval functionality, VDA BV (Hilversum) building the video capturing software, the University of Twente and the LT Lab of DFKI GmbH Saarbrücken, responsible among others for the natural language technology, LIMSI-CNRS (Orsay, France) and Vecsys SA (Les Ulis, France) developing and integrating the speech recognition modules, respectively.

OLIVE is making use of speech recognition in English, French and German to automatically derive transcriptions of the sound tracks, generating time-coded linguistic elements which serve as the basis for text-based retrieval functionality. Confidence scores are associated with each hypothesized word to allow further processing steps to take into account the reliability of the candidates.

Taking advantage of the corpora available through the LDC, the speech recognizer[18, 21] has been developed and tested on American English. The acoustic models are trained on 150 hours of transcribed audio data, with the language models trained on 200M words broadcast news transcriptions and 400M words of newspaper and newswire texts. Using broadcast data collected in OLIVE, LIMSI has ported its American English system to French. A port to German is underway.

Experiments with 700 hours of unrestricted broadcast news data indicate that word error rates around 20% are obtained for American English. Preliminary experiments in French and German indicate that the word error rates are higher, which can be expected as these languages are more highly inflected than English, and less training data are available. However, it has to be kept in mind, that for the purpose of indexing and retrieval a 100% recognition rate is not necessary, since not every word will have to make it into the index, and not every expression in the index is likely to be queried. Research into the differences between text retrieval and spoken document retrieval indicates that recognition errors do not add new problems for the retrieval task[28].

The broadcast transcription testbed is particularly rich in varying acoustic conditions, topics, domains and languages, with native and non-native speakers. Significant progress in multilingual interoperability can be expected from research in broadcast transcriptions.

4.4. Portability

Porting a speech recognizer to a new language consists mainly in the creation of the language specific acoustic models, pronunciation lexica and language models. As mentioned before the acoustic parameter extraction, the model estimation techniques and the search engine may be considered as language-independent. Porting can thus appear as a rather straightforward process, provided there are sufficient speech and text databases available, together with either a pronunciation lexicon or appropriate letter to sound rules for the pronunciation generation. In the previously described SQALE and CALLHOME programs multilingual resources were provided to the different participants for system development. Porting efforts can then be limited in time to a several months span. Much of the demonstrated progress in speech recognition and spoken language understanding over recent years has been fostered by the availability of large commonly used corpora for system training and evaluation in different languages.

But these resources, while in constant increase are still lacking for many human languages. Especially in military and intelligence applications, interest in exotic languages may arise suddenly and the porting phase should span the shortest duration possible.

4.4.1. Porting using language-dependent resources

In the following we relate some of our experience from the SQALE project where our read speech recognition systems of American English and French have been ported to British English and to German. Language-dependent resources (transcribed speech, text material and pronunciation dictionaries) were available to all partners.

For German the acoustic models were bootstrapped using a mix of French and English models. German acoustic models were then estimated from the PHONDAT read speech database, available for research purposes from the University of Munich. Phondat contains a variety of prompt types including phonetically balanced sentences, a few short stories, isolated letters and train timetable queries. There are a total of 15,000 sentences from 155 speakers. Vocabulary items are rather limited, with only about 1700 different words and the prompt texts are quite different in style from the language model training material (taken from newspaper texts). Despite these relatively mismatched acoustic data as compared to the read newspaper task, and despite the limited amount of distinct lexical items, good recognition performance could be observed for German. But we have to recall two important facts: first the German system used a 65k vocabulary to get acceptable lexical coverage, whereas for the other languages the systems were still using 20k vocabularies. Second the SQALE test sets were designed to achieve similar OOV% rates of about 2% for all languages: the OOV rate with a 20k lexicon without OOV control on the test is 10% in German (2.5% in American English). The OOV problem could be reduced by decompounding compound words, as was done for the numbers during text normalization. Decompounding is however a non-trivial task requiring a refined morphological analysis and

even sometimes semantic information. Many compounds can result in two and more items depending on the degree of morphological analysis carried out. For example consider the following compound word occurring in the training texts: *Bundesbahnoberamtsrat* (approximate translation: *Federal-Rail-Head-Office-Chief*). The following decompositions are possible and semantically correct:
Bundesbahnoberamtsrat → *Bundes Bahn Ober Amts Rat*
Bundesbahnoberamtsrat → *Bundesbahn Ober Amtsrat*
Bundesbahnoberamtsrat → *Bundesbahn Oberamtsrat*
Other decompositions such as:
Bundesbahnoberamtsrat → *Bundes Bahnober Amtsrat*
are possible, but semantically poor. This example clearly illustrates that word compounding in German constitutes an OOV-source, as long the recognition system considers a word to be an item occurring between two spaces.

German system development would have taken benefit from a reliable morphological analyzer, both for the quality of the vocabulary (better coverage) and for the LM (more data to estimate Ngrams). As mentioned before even the pronunciations could have been improved, as a lack of consistency may occur when a given morpheme is observed in a long list of compounds.

To conclude here we can say that porting to a new language can be very fast if all resources are available. A baseline system can then be produced in a short delay. In a second step developments can be carried out to better account for language-specificities: typical pronunciation variants, regional accents, stemming, decompounding for agglutinative languages..., Here years can be spent to move away from a baseline performance.

4.4.2. *Lacking training data for the new language: cross-lingual approaches*

A tentative definition of cross-lingual modeling can be the following: resources from one or multiple source languages are used to estimate models for a new target language. Cross-lingual approaches can apply for acoustic phone modeling as similar sounds are often shared across different languages. A relatively large number of research actions aim at defining multilingual or language-independent acoustic model sets [47, 32, 31]. The availability of language-independent acoustic models reduce the problem of lacking acoustic data in the target language.

For lexical and language modeling however language-dependent resources remain mandatory, at least at the present state-of-art. Progress may be achieved through research areas comprising machine translation, multilingual indexing, speech understanding.

The problem of insufficient training material is addressed in [55]. According to this author the dominant factor with respect to performance is the amount of training data available. The author proposes to use the automatically transcribed test data of the new language to adapt the acoustic models to the new language. The proposed method shows a slight but consistent gain in word accuracy when using a subset of automatically transcribed data, selected using a confidence measure criterion, to adapt acoustic and language models.

5. CONCLUSION

We can consider that present recognition systems are potentially multilingual, as the same family of methods and algorithms apply for developing recognizers in a large variety of languages.

Depending on the level of spoken language representation, a more or less important language-dependency is observed. Whereas the acoustic parameter front-end can be considered as mostly language-independent, words and their pronunciations are completely language-dependent. Successful porting to a new target language then requires appropriate language-specific resources, among the most important are text material and pronunciation lexica. The availability and size of these resources is significantly linked to the final recognizer's performance. Developing multilingual resources is expensive, even if dedicated tools exist and speed up the transcription and annotation process. Porting an ASR system to a new target language requires as minimum resource text material for language modeling and pronunciations for the vocabulary. Baseline performance can then be improved either by increasing the volume of training material and/or by adding language-specific knowledge in the various components [52]. Cross-domain research remains an important area, to ensure reusability of these resources when moving to new application domains and to increase ASR interoperability. To overcome the problem of insufficient or missing data researchers are developing interpolation methods to combine corpora. Language specificities, when accounted for properly, will contribute to optimize the recognizer's performance for the new language.

Other research directions concern more language-independent approaches for speech recognition, and more specifically for acoustic modeling. The IPA phone symbol set can theoretically be used to train a collection of language-independent acoustic phone models covering all possible sounds. Language-independent approaches are being investigated [47, 32, 31], and have shown a certain success in porting systems to new languages. Language-independent models have proven useful in bootstrapping recognizers for a new language. Comparative studies show that a small corpus of language-specific acoustic data (1 hour) then rapidly allows to train or adapt better acoustic models [31].

Lexical modeling comprising the definition of the recognizer's vocabulary (word list) with corresponding pronunciations rely on completely language-dependent resources. Vocabularies are often chosen as frequent words occurring in training text corpora which also ensure a good coverage of the application. To overcome a lack of target text corpora for vocabulary definition, bilingual (multilingual) dictionaries can contribute to port vocabularies from source to target languages. But language-dependent resources are necessary for word level modeling (target language text corpora or multilingual dictionaries, letter to sound rules ...). Statistical language modeling for a new target language generally requires huge amounts of text corpora. New challenging research directions joining the domains of machine translation and cross-language information retrieval may contribute in increasing multilingual interoperability in the future.

Multilingual interoperability in automatic speech recognition can be seen as a goal, as a guiding principle to orient

research away from purely language-dependent towards more language-independent questions. This is an important goal to strive for. As the number of written languages remains relatively low, we can imagine having baseline resources available for a large proportion of written languages in a near future. An important research issue then consists in defining and developing these resources and generic corpora, which allow for easy adaptation across domains and languages. The availability of these resources for a large proportion of the spoken/written languages will allow to judge the multilingual capabilities of present speech recognition technology. As underlined by V. Zue in his keynote paper of Eurospeech'97 [57], real deployment of spoken language technology cannot take place without adequately addressing this problem of portability.

REFERENCES

[1] G. Adda, M. Adda-Decker, J.L. Gauvain, L.F. Lamel "Text Normalization and Speech Recognition in French", Proceedings of the European Conference on Speech Technology, EuroSpeech, Rhodos, September 1997.

[2] M. Adda-Decker, G. Adda, L. Lamel, J.L. Gauvain, "Developments in Large Vocabulary, Continuous Speech Recognition of German," *IEEE-ICASSP-96*, Atlanta 1996.

[3] M. Adda-Decker, L.F. Lamel, J.-L. Gauvain, G. Adda, *"Activities in Multilingual Speech Recognition at LIMSI"*, CRIM/FORWISS Workshop on Progress and Propects of Speech Research and Technology, Montréal, Oct. 1996.

[4] M. Adda-Decker, G. Adda, J.L. Gauvain, L. Lamel, "Design considerations for LVCSR in French," *IEEE ICASSP-99*, Phoenix mars 1999.

[5] S. Armstrong et al., "Multilingual Corpora for Cooperation", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 975-980, May 1998.

[6] J. Baker, "The Dragon System – An Overview," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol ASSP-23, pp. 24-29, Feb. 1975.

[7] C. Barras et al., "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", Proc. 1st International Conference on Language Resources and Evaluation, Granada, May 1998.

[8] K. Berkling, M. Zissmann, "Improving accent identification through knowledge of English syllable structure", *Proc. ICSLP-98*, pp. 89-92, vol. II, Sidney, Dec. 1998.

[9] J. Billa et al., "Multilingual Speech Recognition: the 1996 Byblos CALLHOME System", *Eurospeech-97*, Rhodos, September 1997.

[10] S. Bird, M. Liberman, "Towards a Formal Framework for Linguistic Annotations", *Proc. ICSLP-98*, pp. 3179-3180, vol. VII, Sidney, December 1998.

[11] B. Byrne et al., "Toward Language-Independent Acoustic Modeling", Summer Research Workshop on Speech and Language, CLSP, John Hopkins University, 1999.

[12] L. Chase, "A review of the American SWITCHBOARD and CALLHOME Speech Recognition Evaluation programs", 1st Int Conf. on Language Resources and Evaluation, Granada, vol II, pp. 789-793, May 1998.

[13] C. Corredor-Ardoy, L. Lamel, M. Adda-Decker, J.L. Gauvain, *"Multilingual Phone Recognition of Spontaneous Telephone Speech," IEEE ICASSP-98*, Seattle, WA, 1998.

[14] C. S. Culhane, "Conversational and Multi-lingual Speech Recognition", *Proc. DARPA Speech Recognition Workshop*, pp. , Arden Conference Center, Harriman, New York 1996.

[15] Ch. Draxler, H. van den Heuvel, H. S. Tropf, "SpeechDat Experiences in Creating Large Multilingual Speech Databases for Teleservices", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 361-370, May 1998.

[16] P. Fung et al., "MAP-based cross-language adaptation augmented by linguistic knowledge: from English to Chinese", *Eurospeech-99*, vol.II, pp.871-874, Budapest, September 1999.

[17] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), pp. 21-37, October 1994.

[18] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop*, Feb. 1997, pp. 56-63.

[19] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News Shows," *Proc. IEEE ICASSP-97*, Munich 1997.

[20] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcription of Broadcast News," *Proc. EuroSpeech'97*, Rhodos, Greece, September 1997.

[21] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP'98*, Sydney, Nov. 1998, pp. 1335-1338.

[22] P. Geutner et al., "Transcribing Multilingual Broadcast News using Hypothesis Driven Lexical Adaptation", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia 1998.

[23] "Cross-Language Information Retrieval", book by G. Grefenstette, Editor, The Kluwer International Series on Information Retrieval, Kluwer Academic Publishers, Dordrecht / Boston / London, 1998.

[24] B. Habert, G. Adda, M. Adda-Decker, P. Boula de Mareüil, S. Ferrari, O. Ferret, G. Illouz, P. Paroubek, *"The need for tokenization evaluation"*, Proc. 1st International Conference on Language Resources and Evaluation, Granada, May 1998.

[25] J. M. Huerta et al., "The Development of the 1997 CMU Spanish Broadcast News Transcription", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia 1998.

[26] N. Ide, "Corpus Encoding Standards: SGML Guidelines for Encoding Linguistic Corpora", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 463-469, May 1998.

[27] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, **64**(4), pp. 532-556, 1976.

[28] G. Jones, J. Foote, K. Sparck Jones and S. Young, "The video mail retrieval project: experiences in retrieving spoken documents," Mark T. Maybury (ed.) *Intelligent Multimedia Information Retrieval*, AAAI Press, 1997.

[29] F.M.G. de Jong "Twenty-One: a baseline for multilingual multimedia retrieval", *Proceedings of the 14th Twente Workshop on Language Technology (TWLT-14)*, University of Twente, 1998, pp. 189-194.

[30] F. de Jong, J.L. Gauvain, J. den Hartog, K. Netter, "OLIVE: Speech Based Video Retrieval", to appear in CBMI'99, European Workshop on Content-Based Multimedia Indexing, Toulouse, France, October 1999.

[31] S. Khudanpur et al., "Cross-Language Adaptation of Acoustic Models", Summer Research Workshop on Speech and Language, CLSP, John Hopkins University, 1999.

[32] Köhler J., "Language-adaptation of multilingual phone models for vocabulary independent speech recognition tasks", *Proc. IEEE ICASSP-98*, pp. 417-420, vol. I, Seattle May 1998. *Eurospeech'95*, Madrid, Sept. 1995.

[33] F. Kubala, "Broadcast News is Good News", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 83-87, Washington 1999.

[34] L. Lamel, J.-L. Gauvain, "Cross-lingual experiments with phone recognition", *IEEE-ICASSP-93*, vol.2, pp. 507-510, April 1993.

[35] L.F. Lamel, M. Adda-Decker, J.L. Gauvain "Issues in Large Vocabulary, Multilingual Speech Recognition," *Eurospeech-95*, Madrid, September 1995.

[36] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J.J. Gangolf, J.L. Gauvain, *"A Multilingual Corpus for Language Identification,"* Proc. 1st International Conference on Language Resources and Evaluation, Granada, May 1998.

[37] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, 9(2), pp. 171-185, 1995.

[38] M. Malherbe, "Les Langages de l'Humanité", Bouquins collection, Robert Laffont editor, 1995 Paris.

[39] J. Mariani, L. Lamel, "An Overview of EU Programs Related to Conversational/Interactive Systems", Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 247-253, Lansdale, February 1998.

[40] J. Mariani, P. Paroubek, Human Language Technologies Evaluation in the European Framework", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 237-242, Herndon, Virginia 1999.

[41] D. Matrouf, M. Adda-Decker, L.F. Lamel, J.L. Gauvain, "Language identification incorporating lexical information," *ICSLP-98*, Sidney, November 1998.

[42] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, K. Shirai, "Large Vocabulary Continuous Speech Recognition using a Japanese Business Newspaper (Nikkei)", *Proc. DARPA Speech Recognition Workshop*, pp.137-142, Arden Conference Center, Harriman, New York 1996.

[43] D. Pallett, "The NIST Role in Automatic Speech Recognition Benchmark Tests", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 327-330, 1998.

[44] Rabiner, L.R. and Juang, B.H.: "An introduction to Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 3(1), pp. 4-16, 1986.

[45] N. Ruimy et al. "The European LE-PAROLE Project: the Italian Syntactic Lexicon", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 241-248, 1998.

[46] T. Schultz, A. Waibel, "Fast Bootstrapping of LVSR Systems with Multilingual Phoneme Sets", *Eurospeech-97*, pp. 371-374, Rhodos, September 1997.

[47] T. Schultz, A. Waibel, "Language-independent and language adaptive large vocabulary speech recognition", *Proc. ICSLP-98*, pp. 1819-1822, vol. V, Sidney, Dec. 1998.

[48] R. Schwartz, H. Jin, F. Kubala, S. Matsoukas, "Modeling Those F-Conditions – Or Not," *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 115-118, February 1997.

[49] "Multilingual Text-to-Speech Synthesis - The Bell Labs Approach", book by R. Sproat, Editor, Kluwer Academic Publishers, Dordrecht / Boston / London, 1998.

[50] R. Stern et al., "Specification for the ARPA November 1996 Hub 4 Evaluation," Nov. 1996. *Proc. DARPA Speech Recognition Workshop*, pp. , Arden Conference Center, Harriman, New York 1996.

[51] D. Tufis, N. Ide, Tomaz Erjavec, "Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages" 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 233-239, May 1998.

[52] U. Uebler, H. Niemann, "Morphological modeling of word classes for language models", *Proc. ICSLP-98*, pp. 1687-1690, vol. V, Sidney, December 1998.

[53] S. Young and G. Bloothooft, Eds., *Corpus-based methods in language and speech processing*, Dordrecht, The Netherlands: Kluwer Academic Publishers, 1997.

[54] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual large vocabulary speech recognition: the European SQALE project," in *Computer Speech and Language*, volume 11, nb.1, January 1997,pages 73-99.

[55] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance" Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 301-305, Lansdale, February 1998.

[56] P. Zhan et al., "Dragon Systems' 1997 Mandarin Broadcast News System", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, 1998.

[57] V. Zue, "Conversational interfaces: advances and challenges", *Eurospeech-97*, vol.I, pp.KN9-17, Rhodos, September 1997 .